

INTRODUCTION TO DATA SCIENCE

CAPSTONE PROJECT

PREDICTING STOCK PRICES WITH GAUSSIAN PROCESS

Nguyễn Viết Mạnh Khoa - 20184278

Phí Hoàng Long - 20184288

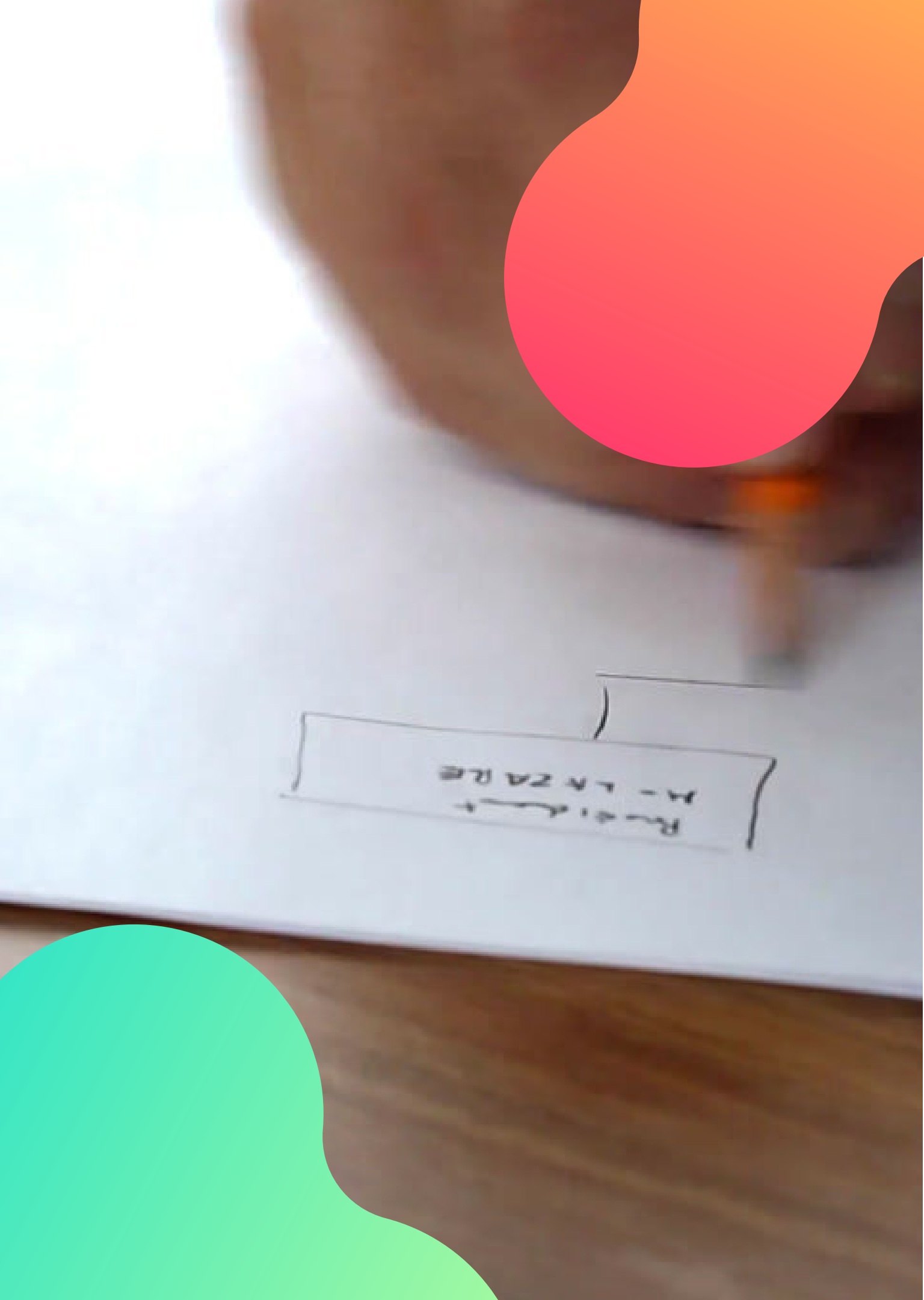
Lê Minh Hiếu - 20184257

OUTLINE

1. Introduction
2. Data crawling
3. Data visualization
4. Algorithm
5. Evaluation

Work Contribution

- Le Minh Hieu: crawling data and preprocessing data
- Phi Hoang Long: visualize data and give insight observations
- Nguyen Viet Manh Khoa: implementing Gaussian Process method to predict stock price and other methods for comparison purpose.



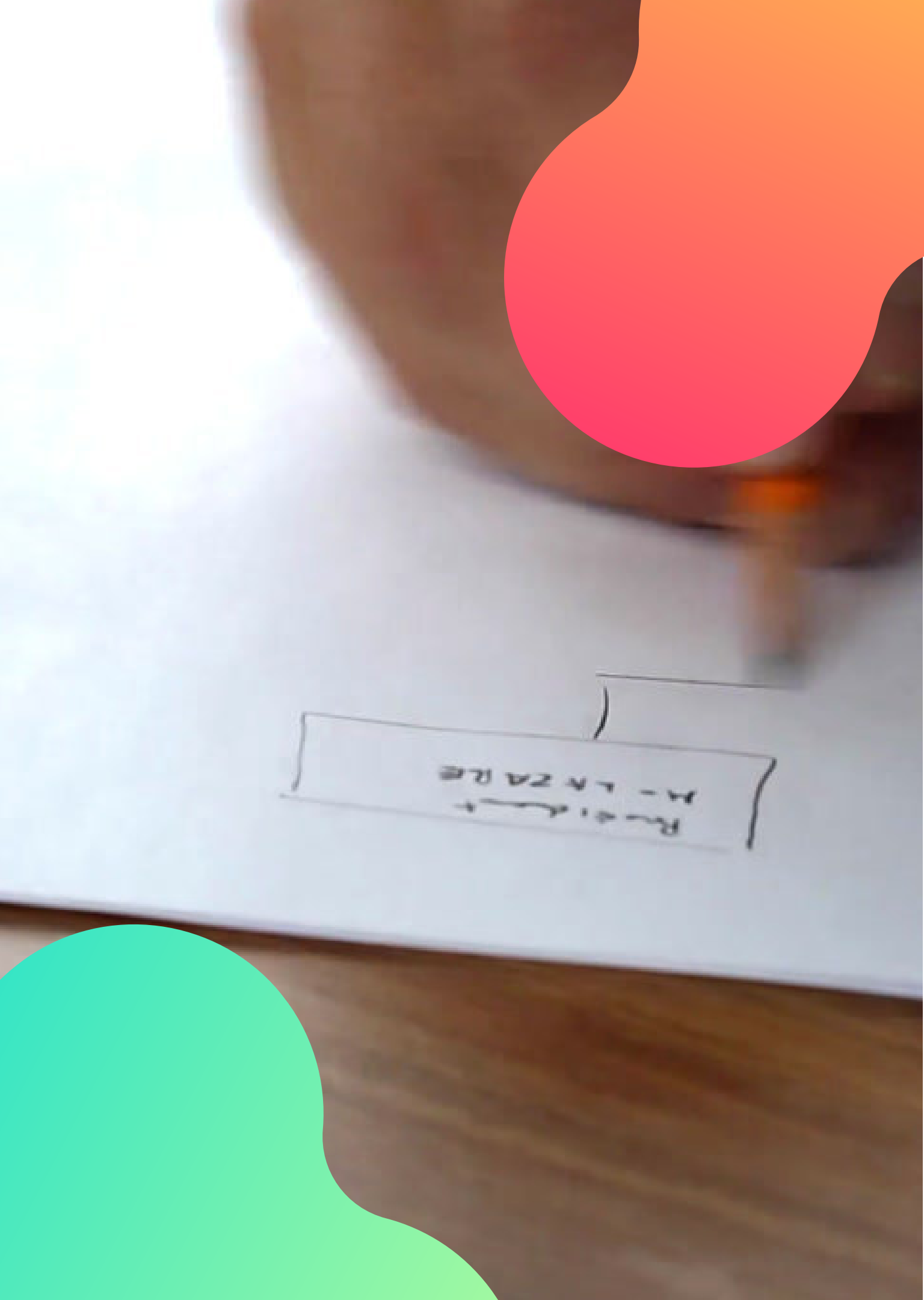
INTRODUCTION

Introduction – Stock price prediction

- A classic and important problem.
- Gain insight about market behavior over time
- Spotting trends hard to predict by human means
- Machine learning is an efficient method

Introduction – Learning problem

- Task: predict and gain future stock price information
- Experience: past stock price trend pattern
- Performance: accuracy with ground truth



DATA CRAWLING

Yahoo Finance

1. NASDAQ Composite (^IXIC)

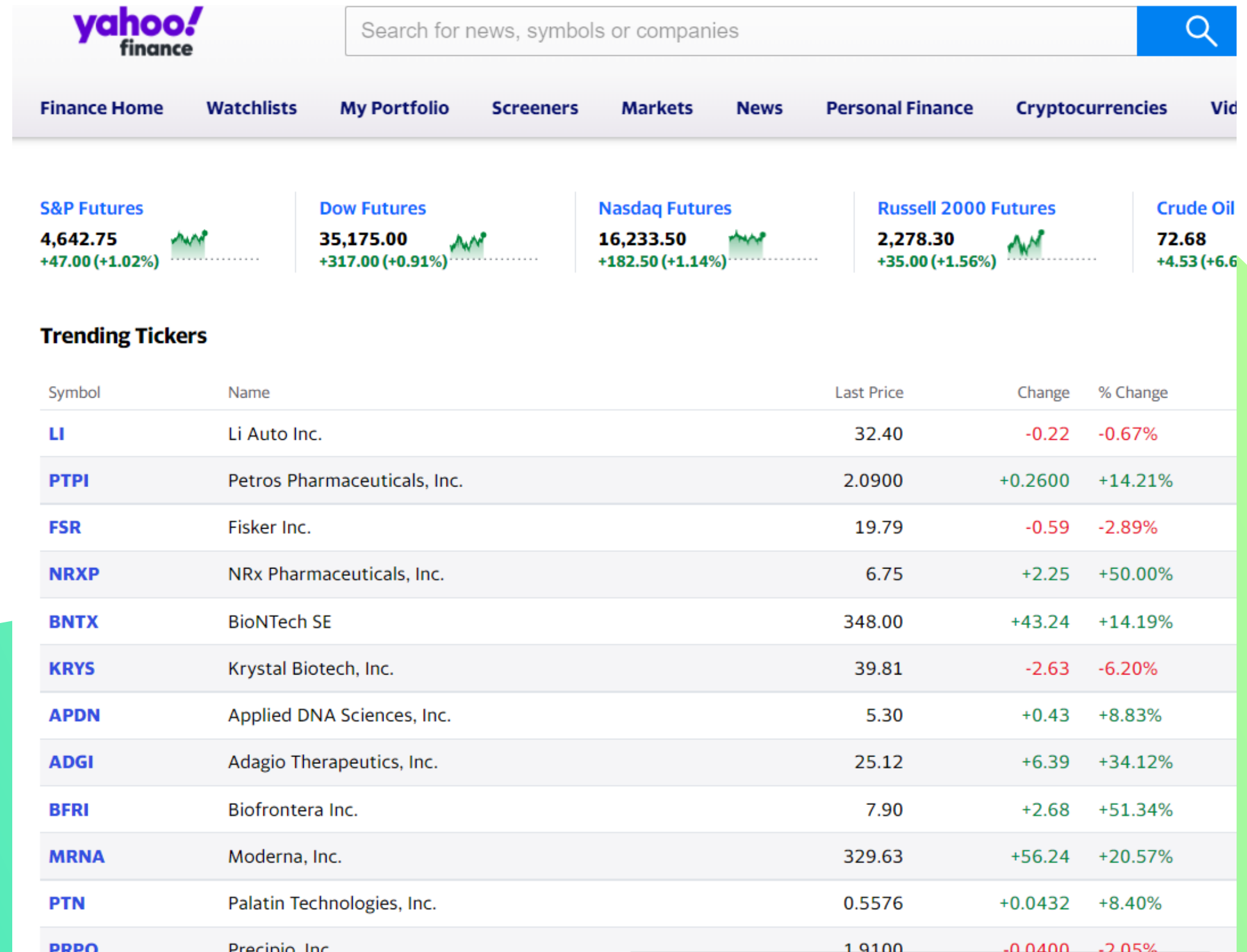
Includes almost all stocks on NASD

2. Palatin Technologies (PTN)

A biopharmaceutical company

3. Cassava Sciences, Inc. (SAVA)

A clinical-stage biotechnology company



Data crawling – Requirements

1 **BeautifulSoup**

Parse HTML/XML
documents

2 **Pandas**

Export files

3 **Selenium**

Controls browser

Data description

1 **Date**

Day of trading

2 **Open**

Starting price in day

3 **Close**

Ending price in day

4 **Adjusted close**

Close price, adjusted after business actions

5 **High**

Highest price in day

6 **Low**

Lowest price in day

7 **Volume**

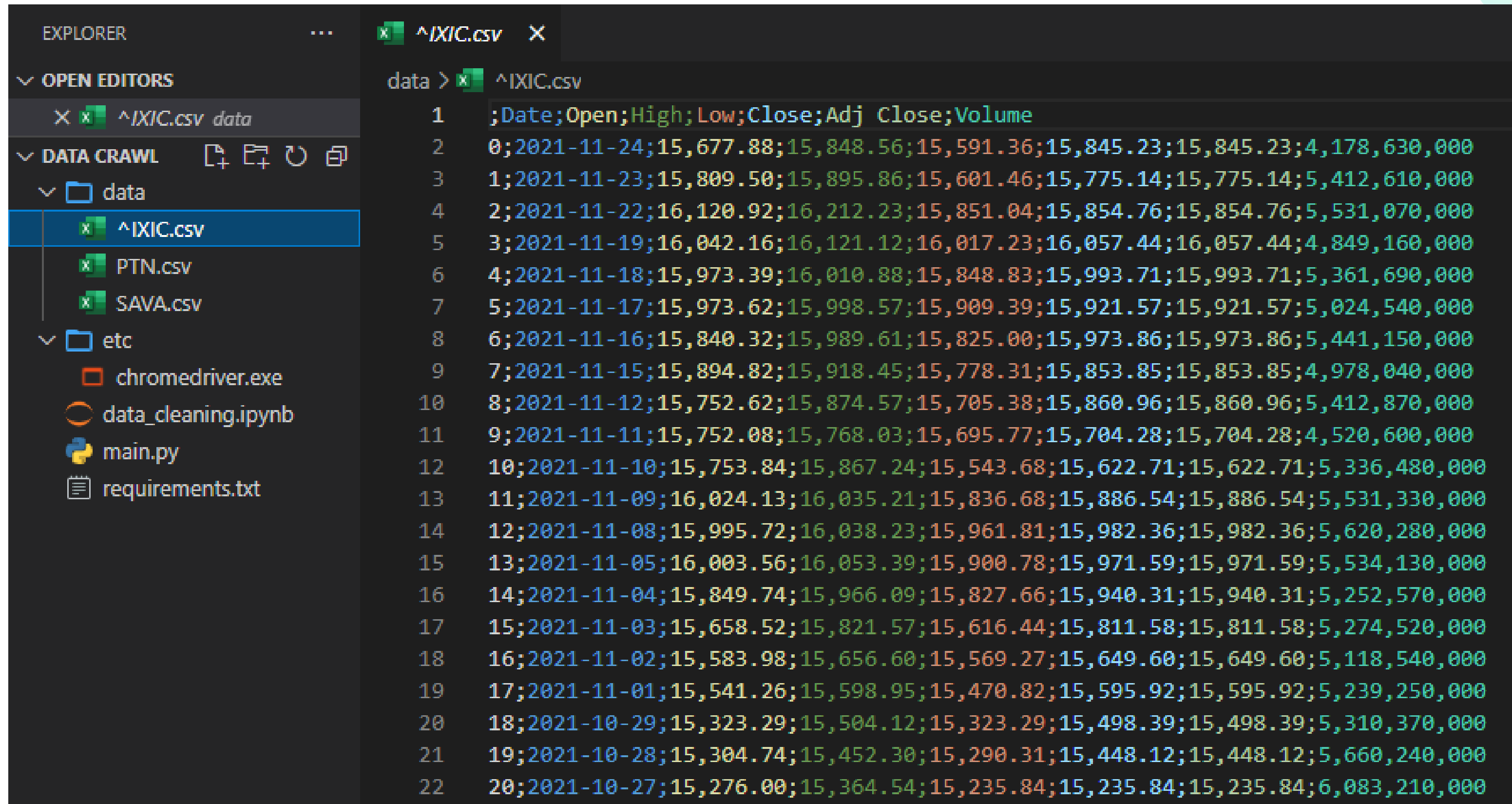
Number of shares traded in day

Data crawling - Steps



1. Use Selenium to drive a browser, to go to source page
2. Scroll to bottom of page until no longer scrollable
3. Extract table rows from complete page
4. Export data to .csv file

Data crawling - Results



EXPLORER

OPEN EDITORS

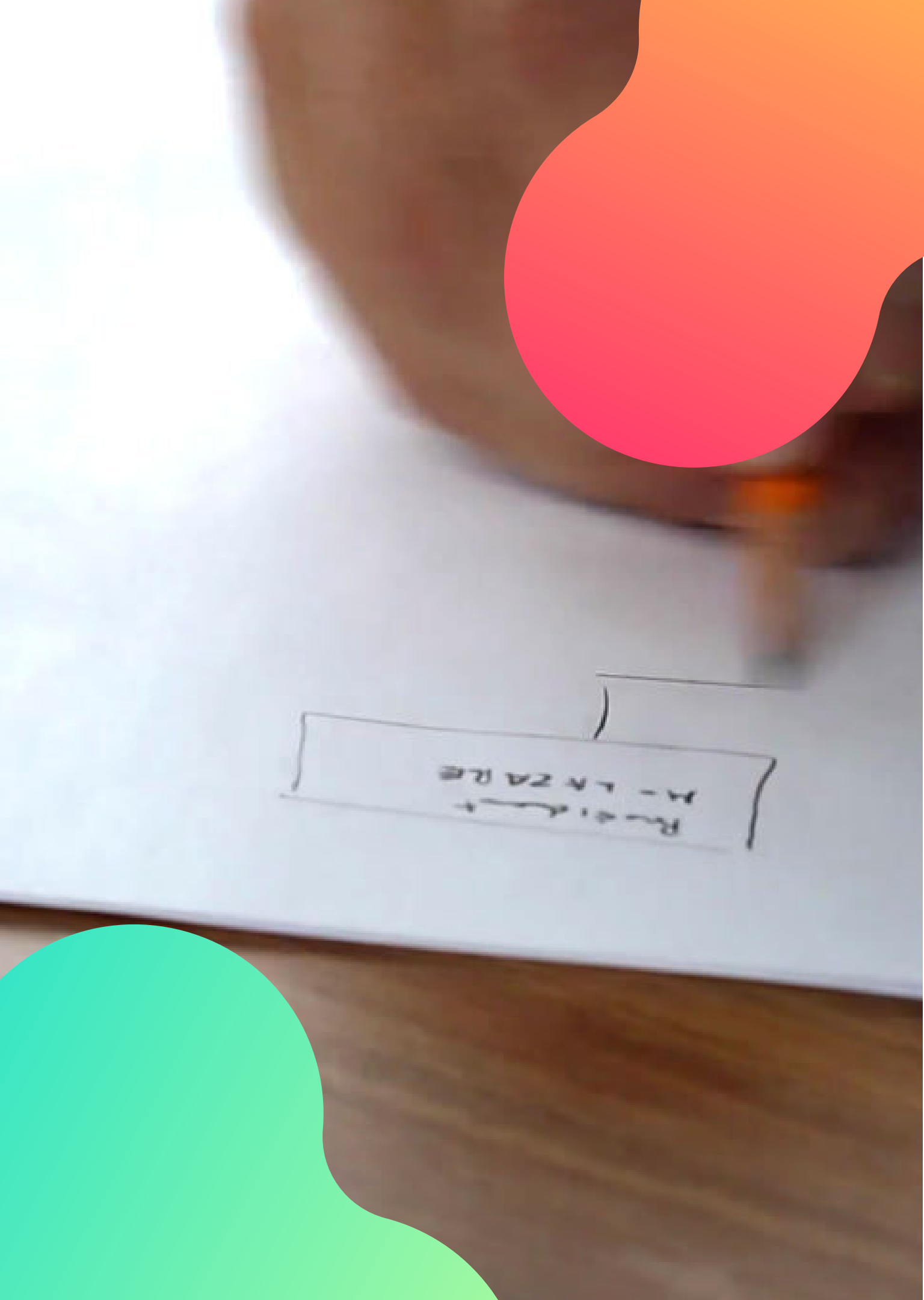
- ^IXIC.csv data

DATA CRAWL

- data
 - ^IXIC.csv
 - PTN.csv
 - SAVA.csv
- etc
 - chromedriver.exe
 - data_cleaning.ipynb
 - main.py
 - requirements.txt

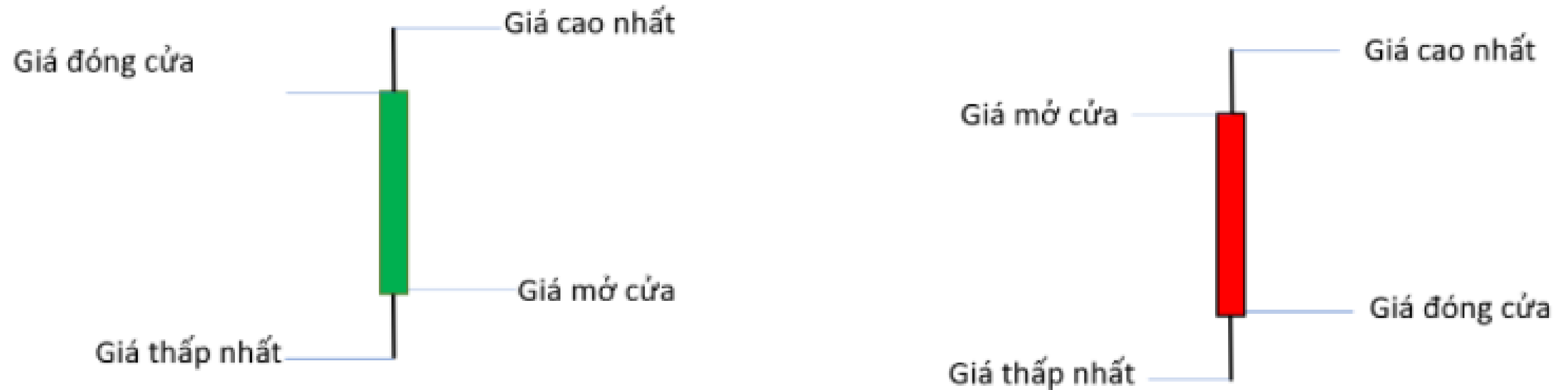
data > ^IXIC.csv

```
1 ;Date;Open;High;Low;Close;Adj Close;Volume
2 0;2021-11-24;15,677.88;15,848.56;15,591.36;15,845.23;15,845.23;4,178,630,000
3 1;2021-11-23;15,809.50;15,895.86;15,601.46;15,775.14;15,775.14;5,412,610,000
4 2;2021-11-22;16,120.92;16,212.23;15,851.04;15,854.76;15,854.76;5,531,070,000
5 3;2021-11-19;16,042.16;16,121.12;16,017.23;16,057.44;16,057.44;4,849,160,000
6 4;2021-11-18;15,973.39;16,010.88;15,848.83;15,993.71;15,993.71;5,361,690,000
7 5;2021-11-17;15,973.62;15,998.57;15,909.39;15,921.57;15,921.57;5,024,540,000
8 6;2021-11-16;15,840.32;15,989.61;15,825.00;15,973.86;15,973.86;5,441,150,000
9 7;2021-11-15;15,894.82;15,918.45;15,778.31;15,853.85;15,853.85;4,978,040,000
10 8;2021-11-12;15,752.62;15,874.57;15,705.38;15,860.96;15,860.96;5,412,870,000
11 9;2021-11-11;15,752.08;15,768.03;15,695.77;15,704.28;15,704.28;4,520,600,000
12 10;2021-11-10;15,753.84;15,867.24;15,543.68;15,622.71;15,622.71;5,336,480,000
13 11;2021-11-09;16,024.13;16,035.21;15,836.68;15,886.54;15,886.54;5,531,330,000
14 12;2021-11-08;15,995.72;16,038.23;15,961.81;15,982.36;15,982.36;5,620,280,000
15 13;2021-11-05;16,003.56;16,053.39;15,900.78;15,971.59;15,971.59;5,534,130,000
16 14;2021-11-04;15,849.74;15,966.09;15,827.66;15,940.31;15,940.31;5,252,570,000
17 15;2021-11-03;15,658.52;15,821.57;15,616.44;15,811.58;15,811.58;5,274,520,000
18 16;2021-11-02;15,583.98;15,656.60;15,569.27;15,649.60;15,649.60;5,118,540,000
19 17;2021-11-01;15,541.26;15,598.95;15,470.82;15,595.92;15,595.92;5,239,250,000
20 18;2021-10-29;15,323.29;15,504.12;15,323.29;15,498.39;15,498.39;5,310,370,000
21 19;2021-10-28;15,304.74;15,452.30;15,290.31;15,448.12;15,448.12;5,660,240,000
22 20;2021-10-27;15,276.00;15,364.54;15,235.84;15,235.84;15,235.84;6,083,210,000
```



DATA VISUALIZATION

Price volatility: Candlestick charts



A candle represents price behavior in a day

- Opening price vs. closing price
- Highest price and lowest price

Trading volume



Number of shares traded over a period of time

Sample candlestick chart analysis



- Case A
- Case B

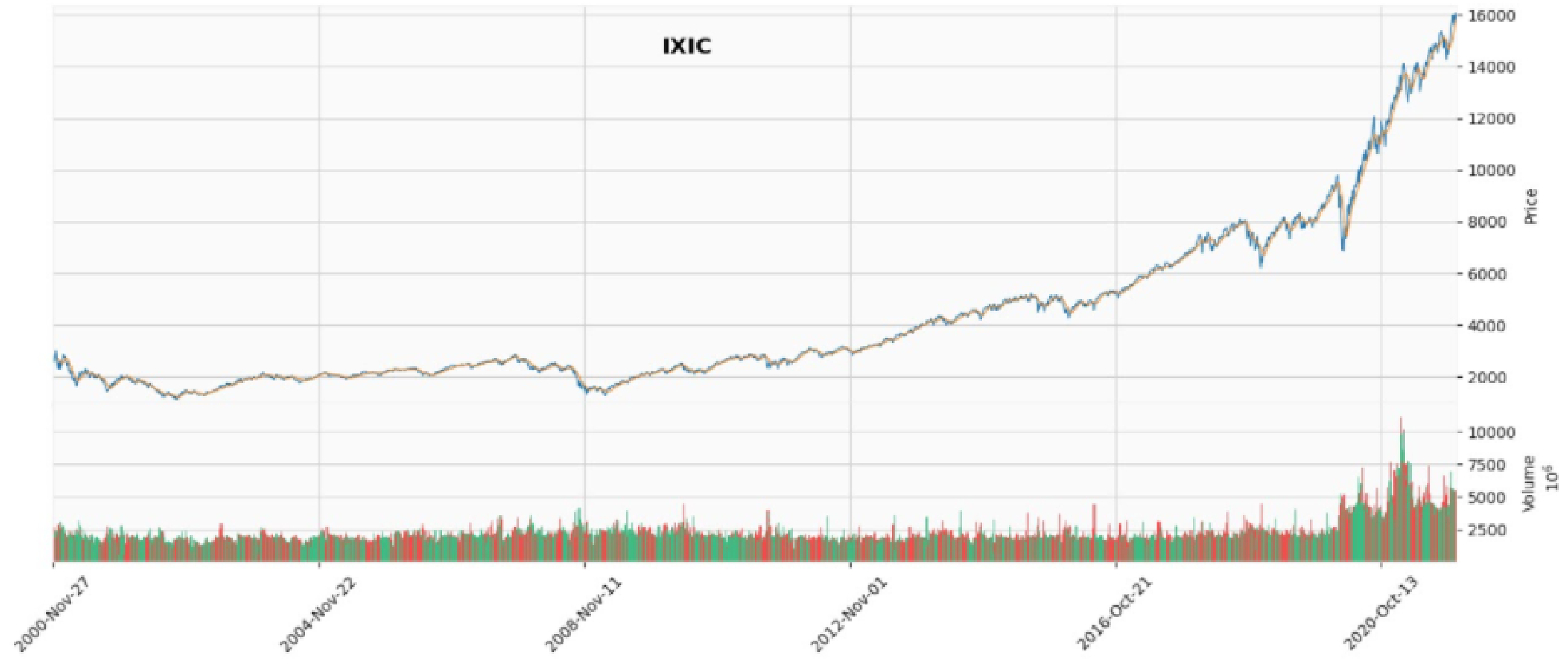


Overall observations (25/11/2000 - 25/11/2021)

- NASDAQ Composite (^IXIC)
 - Palatin Technologies (PTN)
 - Cassava Sciences, Inc. (SAVA)
- 

NASDAQ Composite (^IXIC)

Overall observation (25/11/2000 - 25/11/2021)



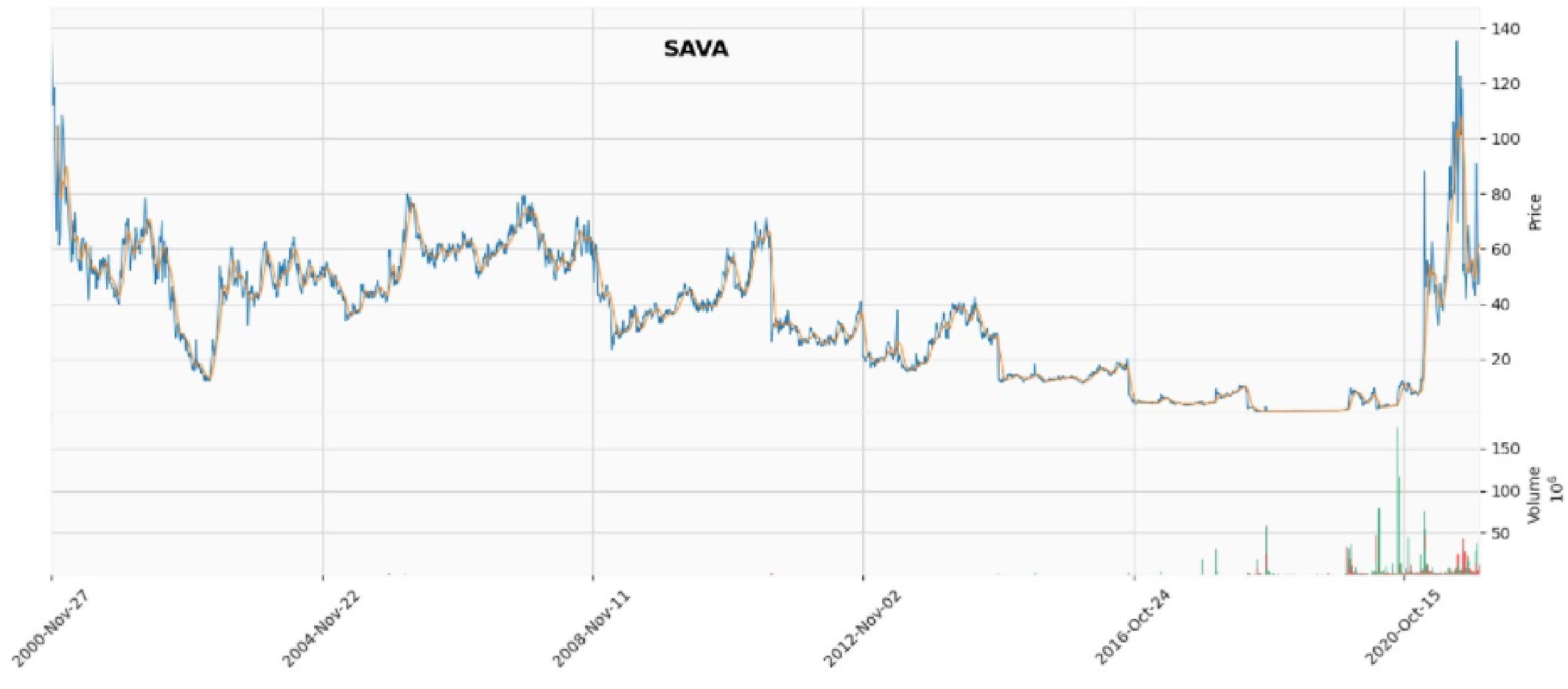
Palatin Technologies

Overall observation (25/11/2000 - 25/11/2021)




Cassava Sciences, Inc.

Overall observation (25/11/2000 - 25/11/2021)





Latest observations (3/5/2021- 25/11/2021)

- NASDAQ Composite (^IXIC)
 - Palatin Technologies (PTN)
 - Cassava Sciences, Inc. (SAVA)
- 

NASDAQ Composite (^IXIC)

Latest observation (3/5/2021 - 25/11/2021)



NASDAQ Composite (^IXIC)

Latest observation (3/5/2021 - 25/11/2021)



Palatin Technologies

Latest observation (3/5/2021 - 25/11/2021)



Palatin Technologies

Latest observation (3/5/2021 - 25/11/2021)



Cassava Sciences, Inc.

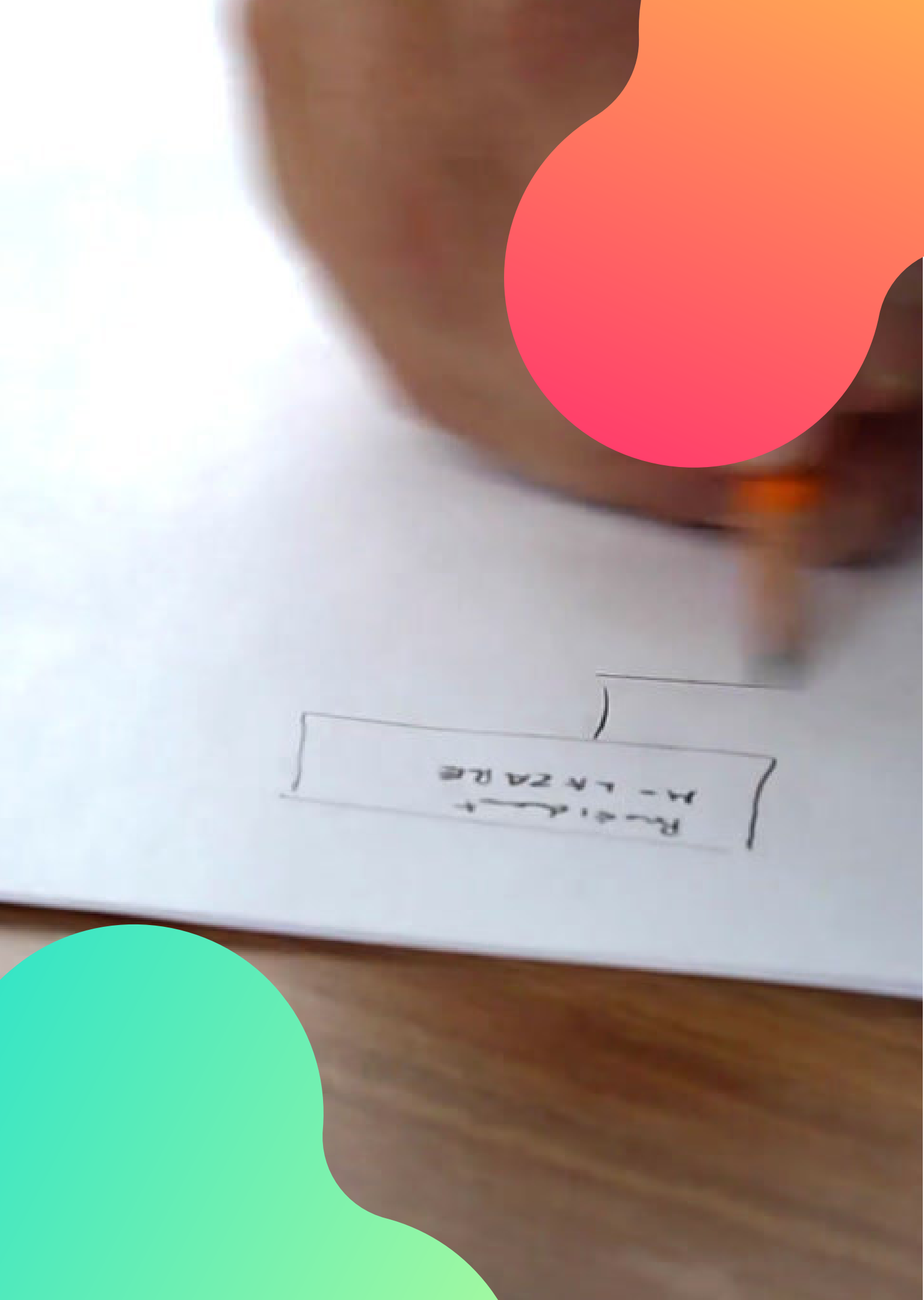
Latest observation (3/5/2021 - 25/11/2021)



Cassava Sciences, Inc.

Latest observation (3/5/2021 - 25/11/2021)





METHOD

Overview: Gaussian Process

A Gaussian process is a generalization of the Gaussian distribution - it represents a probability distribution over *functions* which is entirely specified by a mean and covariance *functions*. Mathematical definition would be then as follows:

Definition: *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Let x be some process $f(x)$. We write:

$$f(x) \sim GP(m(\cdot), k(\cdot, \cdot)),$$

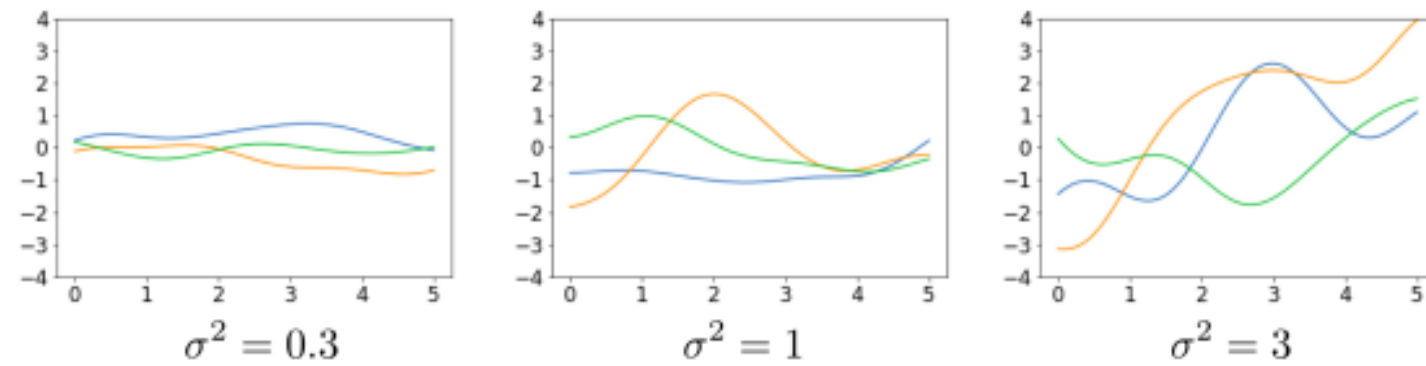
where $m(\cdot)$ and $k(\cdot, \cdot)$ are the mean and covariance functions, respectively:

$$\begin{aligned} m(x) &= E[f(x)] \\ k(x_1, x_2) &= E[(f(x_1) - m(x_1))(f(x_2) - m(x_2))]. \end{aligned}$$

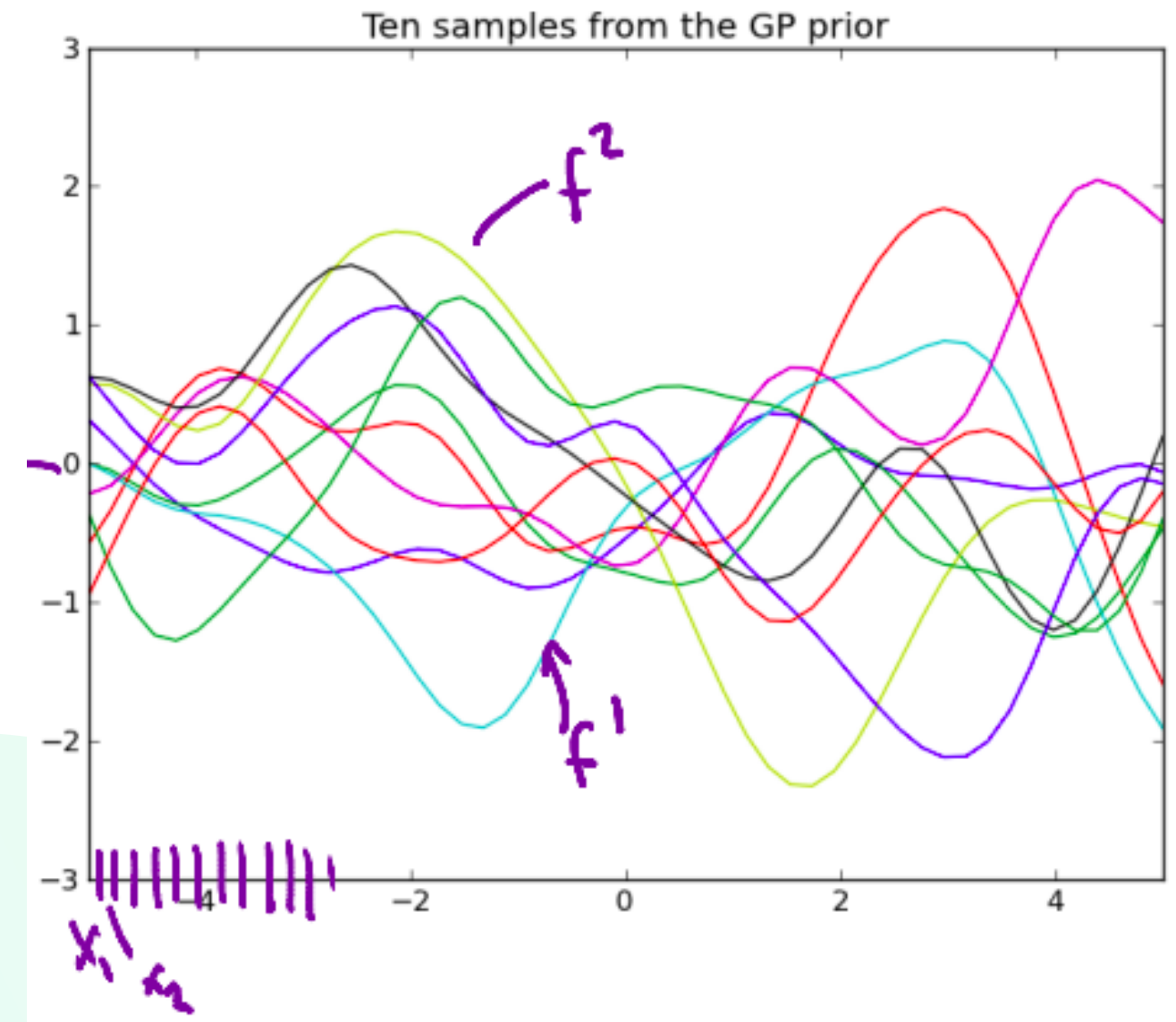
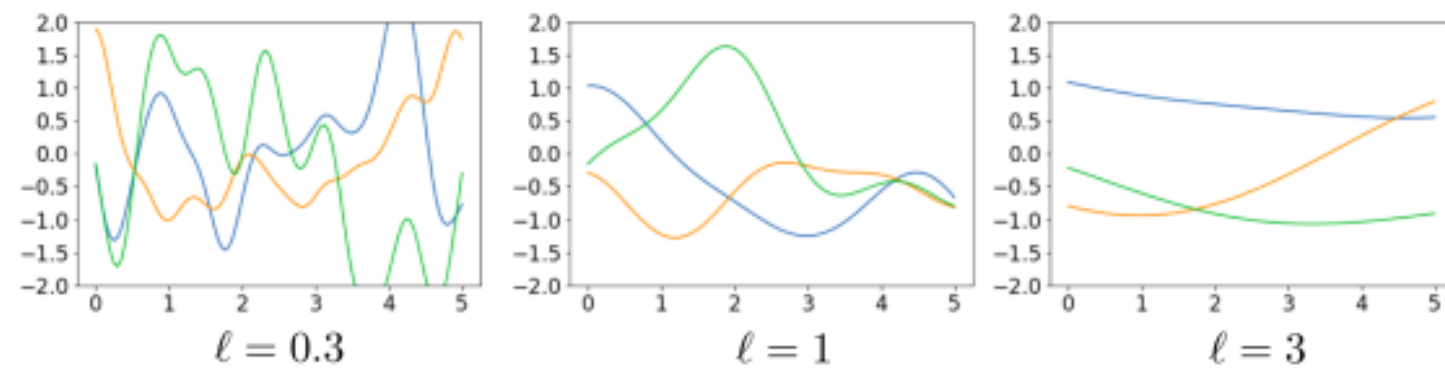
Hyperparameters: GP Kernel

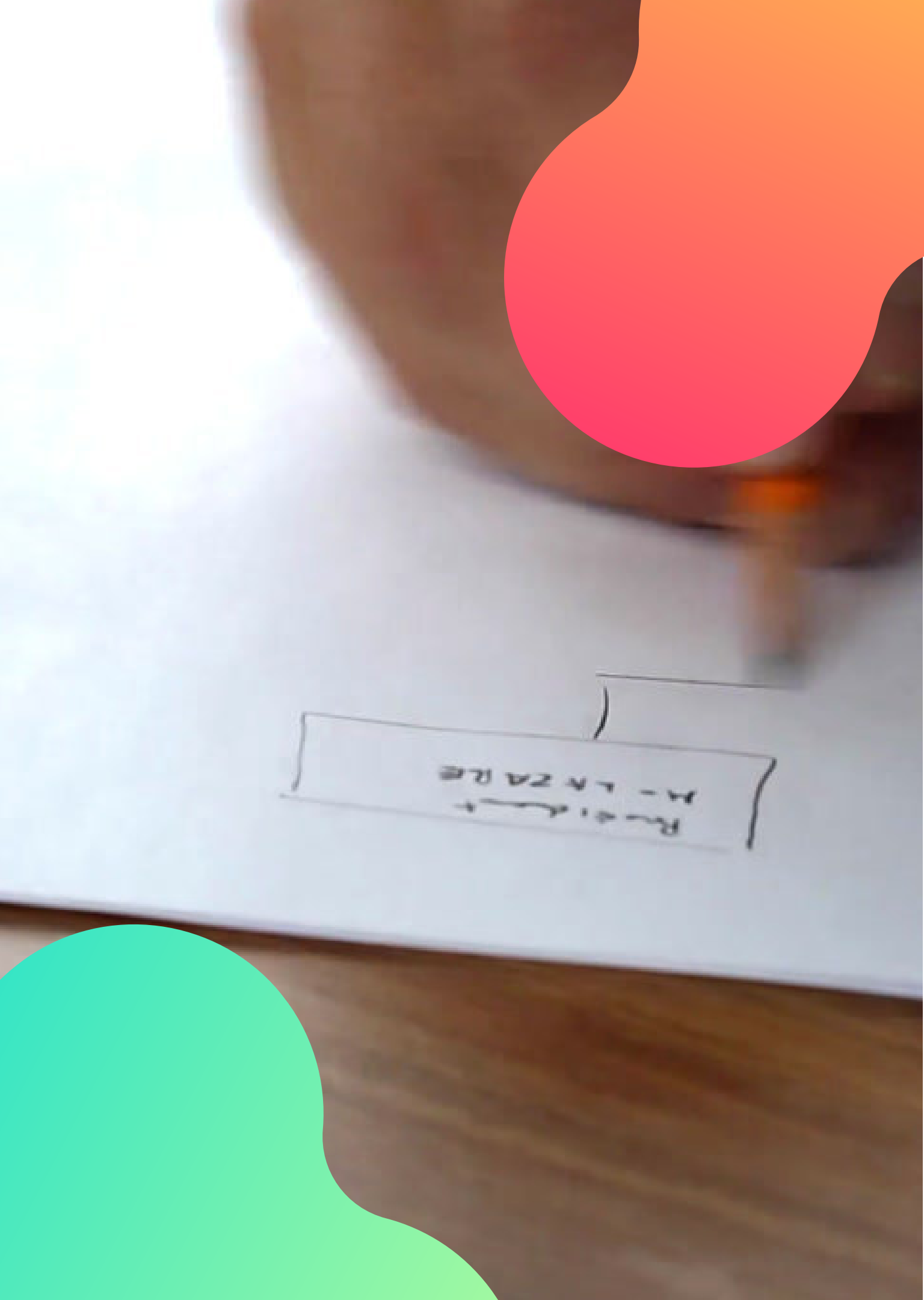
$$k_{\text{SE}}(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right)$$

- The hyperparameters determine key properties of the function.
- Varying the **output variance** σ^2 :



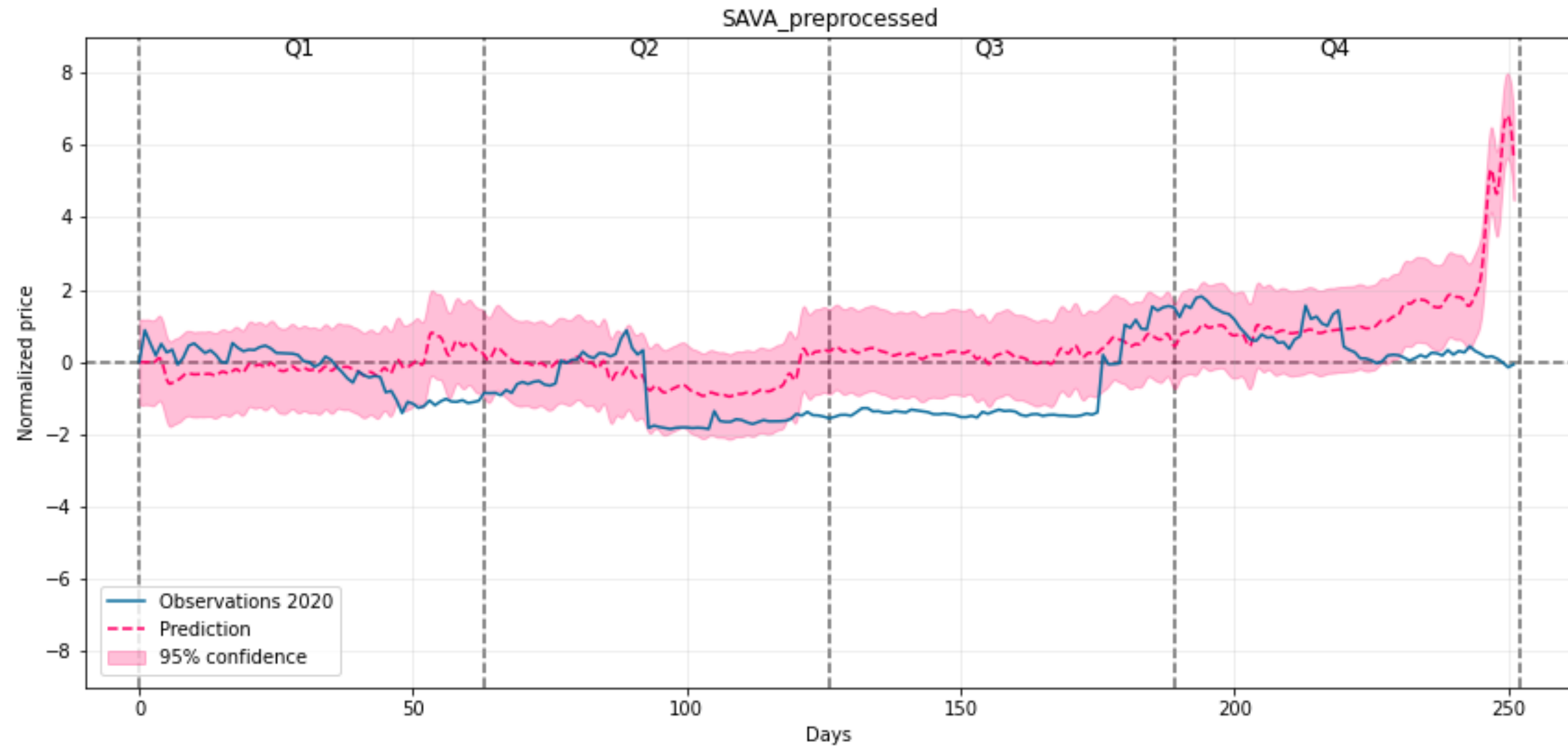
- Varying the **lengthscale** ℓ :



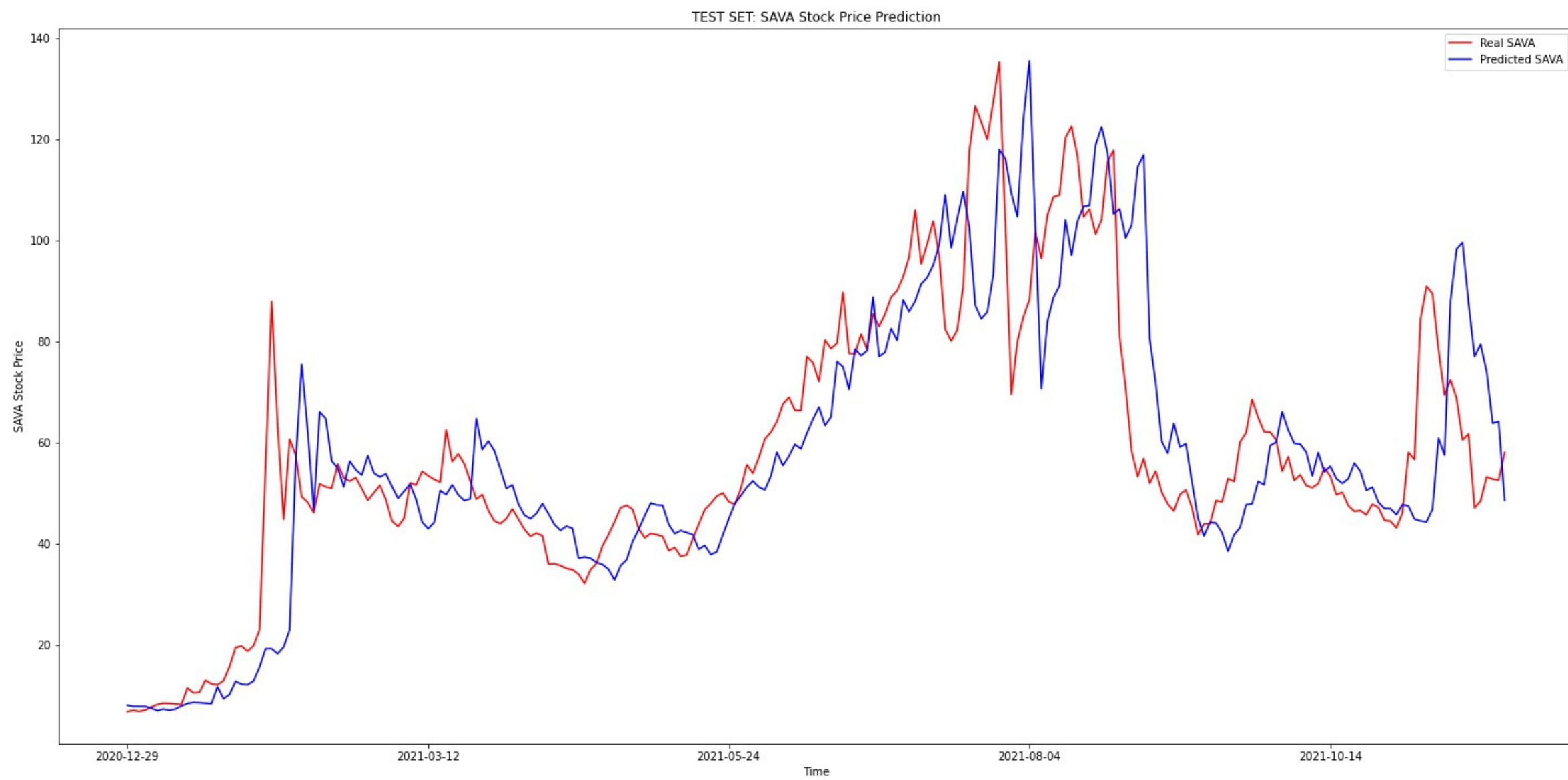


EVALUATION

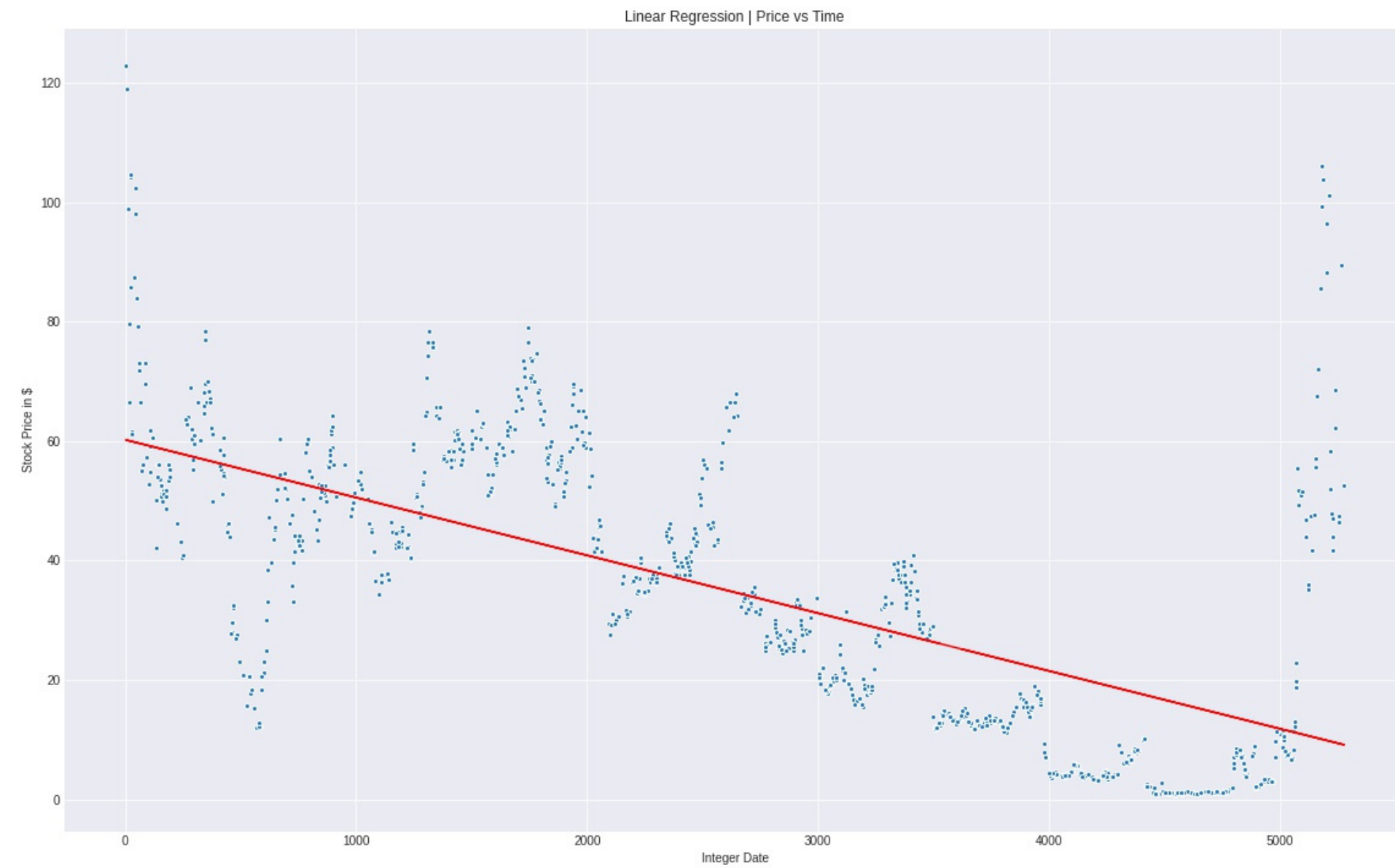
Result on Gaussian Process: SAVA



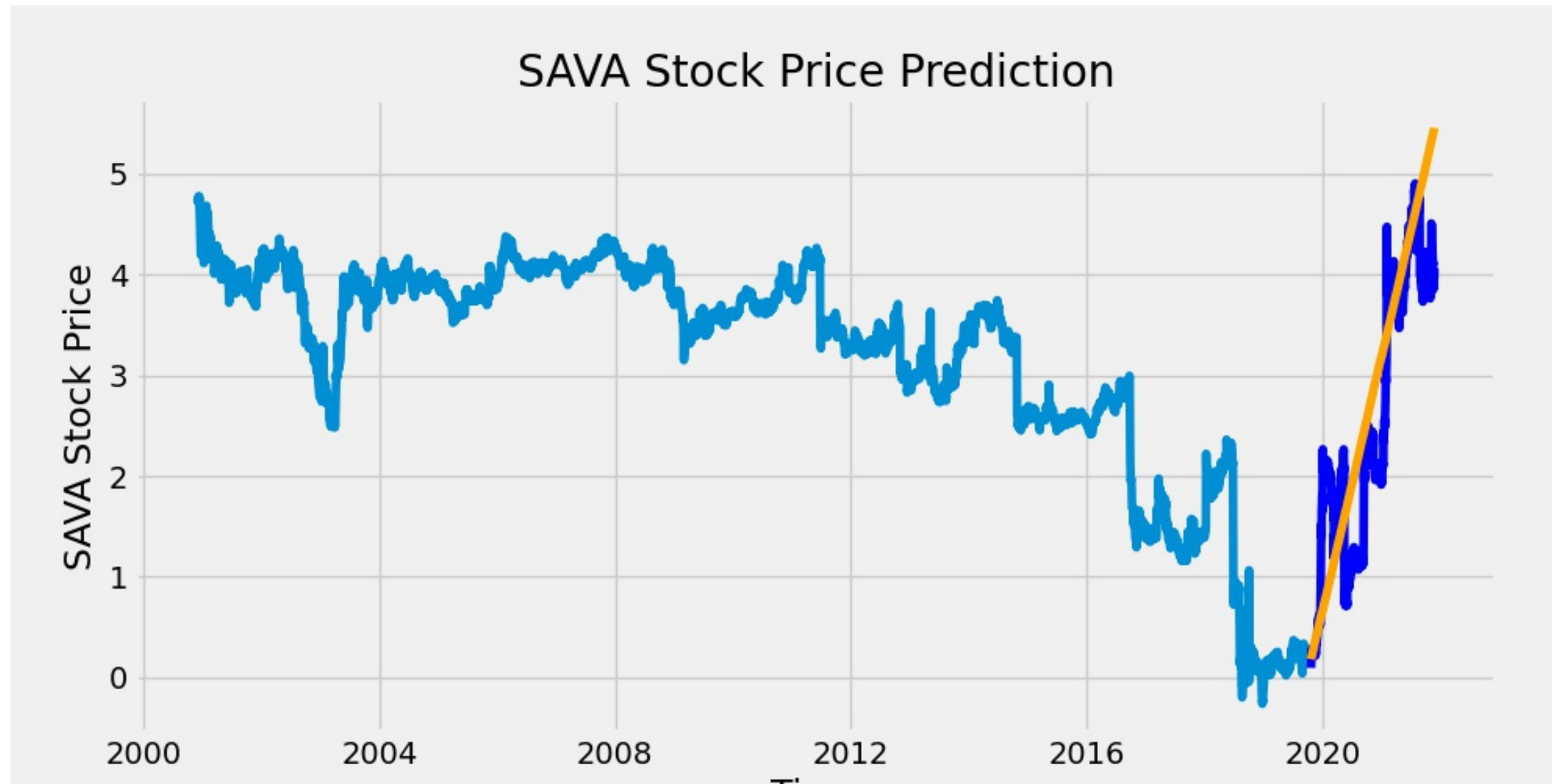
Result on other methods: LSTM




Result on other methods: Linear Regression



Result on other methods: ARIMA



Comparison

Type	Gaussian Process	LSTM	Linear Regression	ARIMA model
MSE	152.5147257	28.7624804	282.5947357	0.56767490829
MAE	8.15730316 	4.172498876	12.29430316	0.610451537752